Instability, Computational Efficiency, and Statistical Accuracy

Nhat Ho

University of Texas, Austin

Joint work with Raaz Dwivedi, Koulik Khamaru, Martin J. Wainwright, Michael I. Jordan, Bin Yu

September, 2021

Talk outline

1. Challenges with optimization methods in parametric statistical models

2. Population to sample analysis framework
   - Contraction of population operator
   - Stability of sample operator

3. Convergence of optimization methods under different settings of operators
   - Stable and fast operators
   - Stable and slow operators
   - Unstable and fast operators
   - Unstable and slow operators

Main story

Unstable optimization algorithms can be preferred to stable algorithms in some statistical settings.

Parametric statistical models

- Given a random sample of size n

$$X_1, \ldots, X_n \sim f_{\theta^\star}(x)$$

- Known: family of distributions $\{f_\theta(x), \theta \in \Theta\}$
- Unknown: $\theta^\star$

Estimation methods

- Standard approaches to estimate $\theta^\star$ include M-estimators, methods of moments, etc.
- Challenge: f_θ is generally non-convex function and optimal solutions from these approaches do not admit closed-forms
- Solution: Optimization algorithms are used to approximate $\theta^\star$

Fundamental questions

- Under what conditions does an optimization algorithm achieve a statistically optimal rate?
- When is an unstable optimization algorithm, such as Newton's method, preferred to a stable algorithm, such as gradient descent method?

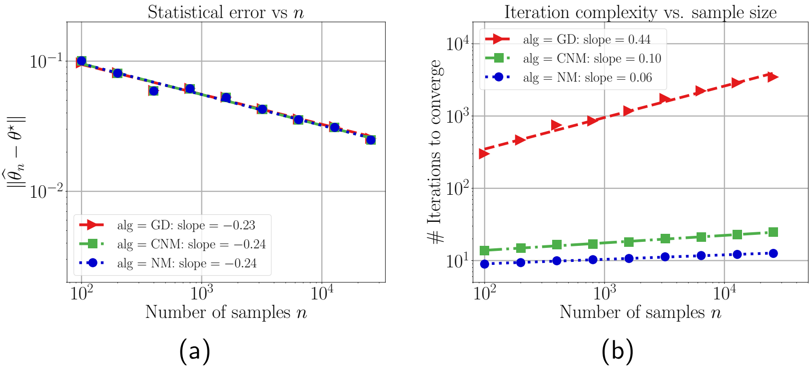
First example: Non-linear regression model

- $\{(X_i, Y_i)\}_{i=1}^n$ are generated from a noisy non-linear regression model of the form

$$Y_i = g\left(X_i^\top \theta^\star\right) + \xi_i, \qquad \text{for } i = 1, \ldots, n.$$

- ξ_i is a zero-mean noise variable with variance σ^2
- $g(t) = t^2$ for $t \in \mathbb{R}$

Behavior of optimization algorithms

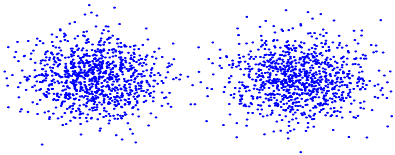


The behavior of gradient descent (GD), cubic-regularized Newton's method (CNM), and the Newton's method (NM) for the regression model when $\theta^\star = 0$.

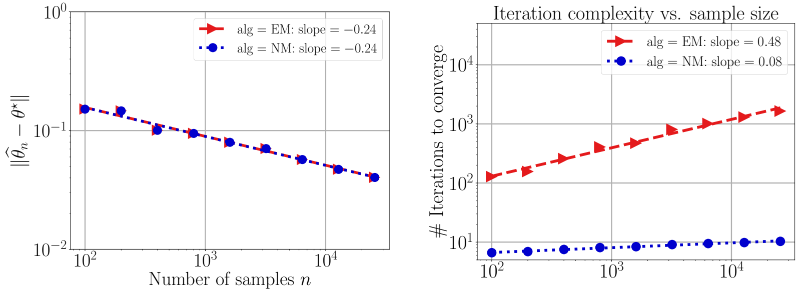
- All the algorithms achieve optimal statistical rates $n^{-1/4}$
- Newton's method takes least number of steps ($\approx \log(n)$) while gradient descent takes significantly larger number of steps ($\approx \sqrt{n}$)

Second example: Mixture model

- Two-component Gaussian mixtures:
  - True model: $\frac{1}{2}\mathcal{N}(-\theta^{\star}, \mathbb{I}_d) + \frac{1}{2}\mathcal{N}(\theta^{\star}, \mathbb{I}_d)$
  - Fitted model: $\frac{1}{2}\mathcal{N}(-\theta, \mathbb{I}_d) + \frac{1}{2}\mathcal{N}(\theta, \mathbb{I}_d)$



Behavior of optimization algorithms



(a) (b)

The behavior of EM algorithm and the Newton's method (NM) for the mixture model when $\theta^\star = 0$.

- EM and Newton's method achieve optimal statistical rates $n^{-1/4}$
- Newton's method takes $\approx \log(n)$ steps to converge while EM algorithm takes significantly larger number of steps ($\approx \sqrt{n}$)

General framework

- F_n: the empirical operator
  - Example: $F_n(\theta) = \theta - \eta \nabla f_n(\theta)$ where f_n is sample log-likelihood function
- F: the population operator
  - Example: $F(\theta) = \theta - \eta \nabla f(\theta)$ where f is population log-likelihood function, i.e., the limit of f_n when $n \to \infty$
- $\theta^\star$: fixed point of F, i.e., $F(\theta^\star) = \theta^\star$
- $\theta_n^{t+1} = F_n(\theta_n^t)$ for $t = 1, 2, \ldots$

Question

Under which conditions, $\{\theta_n^t\}$ approaches a suitably defined neighborhood of $\theta^\star$?

Population to sample analysis

- Triangle inequality:

$$\|\theta_n^{t+1} - \theta^\star\| = \|F_n(\theta_n^t) - \theta^\star\| \leq \underbrace{\|F(\theta_n^t) - \theta^\star\|}_{A} + \underbrace{\|F_n(\theta_n^t) - F(\theta_n^t)\|}_{B}$$

- A: Contraction of population operator

- B: Deviation between sample and population operators

Contraction of population operator F

There are two types of contractions:

- Fast convergence: For $\kappa \in (0,1)$, F is $\texttt{FAST}(\kappa)$-convergent if

$$\|F^t(\theta_0) - \theta^\star\| \leq \kappa^t \|\theta_0 - \theta^\star\| \quad \text{for all } t = 1, 2, \ldots$$

- Slow convergence: For $\beta > 0$, F is $\texttt{SLOW}(\beta)$-convergent if

$$\|F^t(\theta_0) - \theta^\star\| \leq \frac{c}{t^\beta} \quad \text{for all } t = 1, 2, \ldots$$

Example: Fast versus slow convergence

- $\min_\theta f(\theta) = \frac{\theta^{2p}}{2p}$ for some $p \geq 1$
- Gradient descent algorithm:

$$F(\theta) = \theta - \eta \nabla f(\theta) \;=\; \theta\big(1 - \eta\theta^{2p-2}\big)$$

- When $p = 1$, F is $\texttt{FAST}(\kappa)$-convergent algorithm with $\kappa = 1 - \eta$
- When $p \geq 2$, F is $\texttt{SLOW}(\beta)$-convergent with $\beta = \frac{1}{2p-2}$

Deviation between sample and population operators

There are two types of deviations:

- Stability condition: For $\gamma \geq 0$, F_n is $\texttt{STA}(\gamma)$-stable with noise $\varepsilon(\cdot)$ if

$$\mathbb{P}\Big[\sup_{\theta \in \mathsf{Ball}(\theta^\star, r)} \|F_n(\theta) - F(\theta)\| \precsim \min \Big\{ r^\gamma \varepsilon(n, \delta), r \Big\} \Big] \geq 1 - \delta$$

  for any $r > 0$

- Instability condition: For $\gamma < 0$, F_n is $\texttt{UNS}(\gamma)$-unstable with noise $\varepsilon(\cdot)$ if

$$\mathbb{P}\left[\sup_{\theta \in \mathsf{Annulus}(\theta^\star, r, \rho_{\text{out}})} \|F_n(\theta) - F(\theta)\| \leq \varepsilon(n, \delta) \max \left\{ \frac{1}{r^{|\gamma|}}, \rho_{\text{out}} \right\} \right] \geq 1 - \delta$$

  for any radius $r \geq \rho_{\text{in}}$.

Example of stable condition

- $\min_\theta f_n(\theta) = \frac{\theta^4}{4} + \frac{w}{2\sqrt{n}}\theta^2$ where $w \sim N(0, \sigma^2)$
- Gradient descent:
  - Sample operator: $F_n(\theta) = \theta\left(1 - \eta\theta^2 - \eta\frac{w}{\sqrt{n}}\right)$
  - Population operator: $F(\theta) = \theta\left(1 - \eta\theta^2\right)$
- With probability $1 - \delta$,

$$|F_n(\theta) - F(\theta)| = \eta|\theta|\frac{|w|}{\sqrt{n}} \precsim |\theta|\sqrt{\frac{\log(1/\delta)}{n}}$$

$\Longrightarrow F_n$ is $\texttt{STA}(\gamma)$-stable with $\gamma = 1$ and noise $\varepsilon(n, \delta) = \sqrt{\log(1/\delta)/n}$

Example of unstable condition

- $\min_\theta f_n(\theta) = \frac{\theta^4}{4} + \frac{w}{2\sqrt{n}}\theta^2$ where $w \sim N(0, \sigma^2)$
- Newton's method:
  - Sample operator: $F_n(\theta) = \theta - \frac{\theta^3 + w\theta/\sqrt{n}}{3\theta^2 + w/\sqrt{n}}$
  - Population operator: $F(\theta) = \theta - \frac{\theta^3}{3\theta^2}$
- With probability $1 - \delta$, when $|\theta| \gtrsim \left(\frac{\log(1/\delta)}{n}\right)^{1/4}$:

$$|F_n(\theta) - F(\theta)| \lesssim \frac{1}{|\theta|}\sqrt{\frac{\log(1/\delta)}{n}}$$

$\Longrightarrow F_n$ is $\texttt{UNS}(\gamma)$-unstable with parameter $\gamma = -1$ and noise $\varepsilon(n, \delta) = \sqrt{\log(1/\delta)/n}$

General theory: Stable and fast operators

- The operator F is $\texttt{FAST}(\kappa)$-convergent
- The empirical operator F_n is $\texttt{STA}(\gamma)$-stable with noise $\varepsilon(n, \delta)$ for some $\gamma \geq 0$

Theorem 1 (Balakrishnan et al., 2017)

Under suitable initialization, the sequence $\theta_n^{t+1} = F_n(\theta_n^t)$ *satisfies*

$$\|\theta_n^t - \theta^\star\| \precsim \varepsilon(n, \delta) \qquad \textit{when } t \succsim \log(1/\varepsilon(n, \delta)).$$

Furthermore, this bound is tight.

Example of stable and fast operators

- $\{(X_i, Y_i)\}_{i=1}^{n}$ are generated from a noisy non-linear regression model of the form

$$Y_i = (X_i \theta^{\star})^2 + \xi_i, \qquad \text{for } i = 1, \ldots, n.$$

  where $|\theta^{\star}| >>> 1$

- $\xi_i \sim \mathcal{N}(0, 1)$ and $X_i \sim \mathcal{N}(0, 1)$
- We use gradient descent method (GD) to the least-squares loss

Example of stable and fast operators

- Population GD operator F^{GD} is $\texttt{FAST}(\frac{1}{2})$-convergent
- Sample GD operator $\mathrm{F}_n^{\mathrm{GD}}$ is $\texttt{STA}(1)$-stable with noise $\varepsilon(n, \delta) = \sqrt{\frac{\log^4(n/\delta)}{n}}$
- Under suitable initialization, the sequence $\theta_n^{t+1} = \mathrm{F}_n^{\mathrm{GD}}(\theta_n^t)$ satisfies

$$|\theta_n^t - \theta^\star| \precsim n^{-1/2} \qquad \text{when } t \succsim \log(n)$$

General theory: Stable and slow operators

- The population operator F is 1-Lipschitz and is $\texttt{SLOW}(\beta)$-convergent
- The empirical operator F_n is $\texttt{STA}(\gamma)$-stable for some $\gamma \in [0, (1+\beta)^{-1})$

Theorem 2

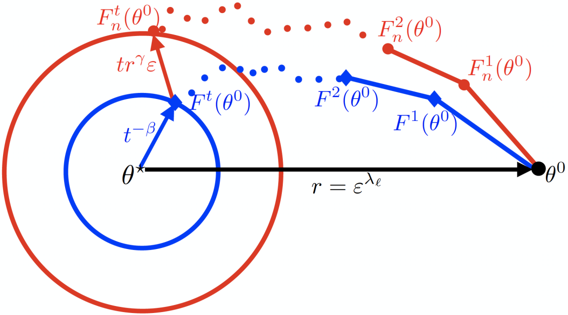
Under suitable initialization, the sequence $\theta_n^{t+1} = F_n(\theta_n^t)$ satisfies

$$\|\theta_n^t - \theta^\star\| \precsim [\varepsilon(n,\delta)]^{\frac{\beta}{1+\beta-\gamma\beta}} \qquad \textit{when } t \succsim \varepsilon(n,\delta)^{-\frac{1}{1+\beta-\gamma\beta}}.$$

Furthermore, this bound is tight.

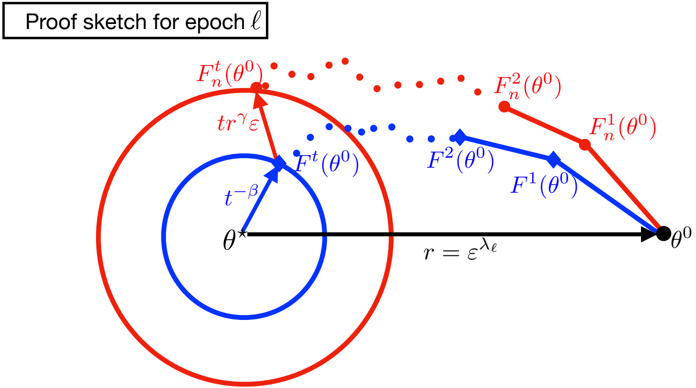
Outline of proof: Epoch-based argument

- Assume θ^0 the starting point for epoch ℓ and $r = \|\theta^\star - \theta^0\| = \varepsilon(n, \delta)^{\lambda_\ell}$
- Slow convergence of population iterates: $\|F^t(\theta^0) - \theta^\star\| \precsim t^{-\beta}$
- Stability of sample operator: $\|F_n^t(\theta^0) - F^t(\theta^0)\| \precsim t \cdot r^\gamma \cdot \varepsilon$



- **Goal:** At the end of epoch ℓ, we want to find suitable t and $\lambda_{\ell+1}$ such that $\|F_n^t(\theta^0) - \theta^\star\| \precsim \varepsilon^{\lambda_{\ell+1}}$

Outline of proof: Epoch-based argument



$$\|F_n^t(\theta^0) - \theta^\star\| \leq \|F_n^t(\theta^0) - F^t(\theta^0)\| + \|F^t(\theta^0) - \theta^\star\| \leq tr^\gamma\varepsilon + \frac{1}{t^\beta} = t\varepsilon^{\gamma\lambda_\ell+1} + \frac{1}{t^\beta} \overset{\text{min over t}}{\asymp} \varepsilon^{\lambda_{\ell+1}}$$

$$\text{where} \quad \lambda_{\ell+1} = \nu\lambda_\ell + \nu' \Longrightarrow \lim_{\ell\to\infty} \lambda_\ell = \nu_\star = \frac{\beta}{1+\beta-\gamma\beta}, \quad \text{and} \quad \nu_\star - \lambda_\ell \leq \alpha \quad \text{for all} \quad \ell \geq \mathcal{O}(\log(1/\alpha))$$

Example of stable and slow operators

- $\{(X_i, Y_i)\}_{i=1}^n$ are generated from a noisy non-linear regression model of the form

$$Y_i = (X_i \theta^\star)^2 + \xi_i, \qquad \text{for } i = 1, \ldots, n$$

  where $\theta^\star = 0$
- $\xi_i \sim \mathcal{N}(0, 1)$ and $X_i \sim \mathcal{N}(0, 1)$
- We apply gradient descent method (GD) to the least-squares loss

Example of stable and slow operators

- Population GD operator:

$$\mathrm{F}^{\mathrm{GD}}(\theta) = \theta\left[1 - 6\eta\theta^2\right]$$

$\Longrightarrow \mathrm{F}^{\mathrm{GD}}$ is $\texttt{SLOW}(\frac{1}{2})$-convergent as $\eta \in (0, \frac{1}{6}]$

- Sample GD operator:

$$\mathrm{F}_n^{\mathrm{GD}}(\theta) = \theta - \eta\left(\frac{2}{n}\sum_{i=1}^{n} X_i^4\theta^3 - \frac{2}{n}\sum_{i=1}^{n} Y_i X_i^2\theta\right)$$

$\Longrightarrow \mathrm{F}_n^{\mathrm{GD}}$ is $\texttt{STA}(1)$-stable with noise $\varepsilon(n,\delta) = \sqrt{\frac{\log^4(n/\delta)}{n}}$

- Under suitable initialization, the sequence $\theta_n^{t+1} = \mathrm{F}_n^{\mathrm{GD}}(\theta_n^t)$ satisfies

$$|\theta_n^t - \theta^\star| \precsim n^{-1/4} \qquad \text{when } t \succsim \sqrt{n}$$

General theory: Unstable and fast operators

- The population operator F is $\texttt{FAST}(\kappa)$-convergent
- The empirical operator F_n is $\texttt{UNS}(\gamma)$-unstable over the annulus $\mathbb{A}(\theta^\star, \widetilde{\rho}_n, \rho)$ for some $\gamma < 0$

Theorem 3

Under suitable initialization, the sequence $\theta_n^{t+1} = F_n(\theta_n^t)$ *satisfies*

$$\min_{k \in \{0,1,\ldots,t\}} \|\theta_n^k - \theta^\star\| \precsim \max\left\{[\varepsilon(n,\delta)]^{\frac{1}{1+|\gamma|}},\ \widetilde{\rho}_n\right\} \qquad \textit{when } t \succsim \log(1/\varepsilon(n,\delta)).$$

Furthermore, this bound is tight.

Outline of proof

- Assume that $\|\theta_n^t - \theta^\star\| > [\varepsilon(n,\delta)]^{\frac{1}{1+|\gamma|}}$ for all $t \precsim \log(1/\varepsilon(n,\delta))$
- As F is $\texttt{FAST}(\kappa)$-convergent and F_n is $\texttt{UNS}(\gamma)$-unstable,

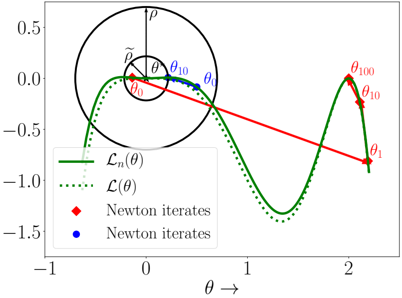
$$\begin{aligned}
\|\theta_n^{t+1} - \theta^\star\| &\leq \|F_n(\theta_n^t) - F(\theta_n^t)\| + \|F(\theta_n^t) - \theta^\star\| \\
&\leq \varepsilon(n,\delta)\max\left\{\frac{1}{[\varepsilon(n,\delta)]^{\frac{|\gamma|}{1+|\gamma|}}}, \rho\right\} + \kappa \cdot \|\theta_n^t - \theta^\star\| \\
&\ldots \\
&\leq \varepsilon(n,\delta)\max\left\{\frac{1}{[\varepsilon(n,\delta)]^{\frac{|\gamma|}{1+|\gamma|}}}, \rho\right\}(1 + \kappa + \ldots + \kappa^{t-1}) \\
&\qquad + \kappa^t \cdot \|\theta_n^0 - \theta^\star\| \\
&\precsim [\varepsilon(n,\delta)]^{\frac{1}{1+|\gamma|}},
\end{aligned}$$

when $t \succsim \log(1/\varepsilon(n,\delta))$

Necessity of the minimum

- We consider the following example:

$$\mathcal{L}(\theta) = -\theta^4(\theta-2)^2 \quad \text{and} \quad \mathcal{L}_n(\theta) = -\left(\theta^4 - \frac{\theta^2}{\sqrt{n}}\right)(\theta-2)^2$$



- When the initialization is too close to $\theta^\star$ (red diamonds), Newton's iterates jump far away from $\theta^\star$ and converge to another fixed point
- When the initialization is in $\mathbb{A}(\theta^\star, \widetilde{\rho}, \rho)$, the Newton iterates (blue circles) do not leave this annulus and converge to a small neighborhood of $\theta^\star$

Additional regularity condition to remove the minimum

- The population operator F is $\texttt{FAST}(\kappa)$-convergent
- The empirical operator F_n is $\texttt{UNS}(\gamma)$-unstable over the annulus $\mathbb{A}(\theta^\star, \widetilde{\rho}_n, \rho)$ for some $\gamma < 0$
- There exists a constant C such that the sequence $\theta_n^t = F_n^t(\theta_n^0)$ satisfies:

$$\|\theta_n^{t+1} - \theta^\star\| \leq C\widetilde{\rho} \quad \text{whenever} \quad \|\theta_n^t - \theta^\star\| \leq \widetilde{\rho},$$

where $\widetilde{\rho} = \max\left\{[\varepsilon(n,\delta)]^{\frac{1}{1+|\gamma|}},\ \widetilde{\rho}_n\right\}$

Proposition 4

Under suitable initialization, the sequence $\theta_n^{t+1} = F_n(\theta_n^t)$ satisfies

$$\|\theta_n^t - \theta^\star\| \precsim \max\left\{[\varepsilon(n,\delta)]^{\frac{1}{1+|\gamma|}},\ \widetilde{\rho}_n\right\} \qquad \textit{when } t \succsim \log(1/\varepsilon(n,\delta)).$$

Furthermore, this bound is tight.

Example of unstable and fast operators

- $\{(X_i, Y_i)\}_{i=1}^n$ are generated from a noisy non-linear regression model of the form

$$Y_i = (X_i \theta^\star)^2 + \xi_i, \qquad \text{for } i = 1, \ldots, n$$

where $\theta^\star = 0$

- $\xi_i \sim \mathcal{N}(0, 1)$ and $X_i \sim \mathcal{N}(0, 1)$
- We apply Newton's method (NM) to the least-squares loss

Example of unstable and slow operators

- Population NM operator:

$$\mathrm{F}^{\mathrm{NM}}(\theta) = \theta - \frac{\theta^3}{3\theta^2} = \frac{2}{3}\theta$$

$\Longrightarrow \mathrm{F}^{\mathrm{NM}}$ is $\texttt{FAST}(\frac{2}{3})$-convergent

- Sample NM operator:

$$\mathrm{F}_n^{\mathrm{GD}}(\theta) = \theta - \frac{\left(\frac{1}{n}\sum_{i=1}^n X_i^4\right)\theta^3 - \left(\frac{1}{n}\sum_{i=1}^n Y_i X_i^2\right)\theta}{\left(\frac{3}{n}\sum_{i=1}^n X_i^4\right)\theta^2 - \frac{1}{n}\sum_{i=1}^n Y_i X_i^2}$$

$\Longrightarrow \mathrm{F}_n^{\mathrm{NM}}$ is $\texttt{UNS}(-1)$-unstable over the annulus $\mathbb{A}(\theta^\star, \widetilde{\rho}_n, 1)$ with $\widetilde{\rho}_n \asymp \log(n/\delta)/n^{1/4}$

Example of unstable and slow operators

- F^{NM} is $\texttt{FAST}(\frac{2}{3})$-convergent
- $\mathrm{F}_n^{\mathrm{NM}}$ is $\texttt{UNS}(-1)$-unstable over the annulus $\mathbb{A}(\theta^\star, \widetilde{\rho}_n, 1)$ with $\widetilde{\rho}_n \asymp \log(n/\delta)/n^{1/4}$
- Additional regularity condition:

$$\left|\mathrm{F}_n^{\mathrm{NM}}(\theta)\right| \geq |\theta_n^*|$$

for all $|\theta| \in [|\theta_n^*|\,, 1]$ where θ_n^* is global solution of least-squares loss

- Under suitable initialization, the sequence $\theta_n^{t+1} = \mathrm{F}_n^{\mathrm{NM}}(\theta_n^t)$ satisfies

$$|\theta_n^t - \theta^\star| \precsim n^{-1/4} \qquad \text{when } t \succsim \log(n)$$

General theory: Unstable and slow operators

- The population operator F is 1-Lipschitz and is $\texttt{SLOW}(\beta)$-convergent
- The empirical operator F_n is $\texttt{UNS}(\gamma)$-unstable over the annulus $\mathbb{A}(\theta^\star, \widetilde{\rho}_n, \rho)$ for some $\gamma < 0$

Theorem 5

Under suitable initialization, the sequence $\theta_n^{t+1} = F_n(\theta_n^t)$ *satisfies*

$$\min_{k \in \{0,1,\ldots,t\}} \|\theta_n^k - \theta^\star\| \precsim \max\left\{ [\varepsilon(n,\delta)]^{\frac{\beta}{1+\beta-\gamma\beta}}, \ \widetilde{\rho}_n \right\} \qquad \textit{when } t \succsim \varepsilon(n,\delta)^{-\frac{1}{1+\beta}}.$$

Furthermore, this bound is tight.

Outline of proof

- $\nu_\star = \frac{\beta}{1+\beta-\gamma\beta}$
- Assume that $\|\theta_n^t - \theta^\star\| > \max\left\{[\varepsilon(n,\delta)]^{\nu_\star},\ \widetilde{\rho}_n\right\}$ for all $t \precsim \varepsilon(n,\delta)^{-\frac{1}{1+\beta}}$
- As F is $\texttt{SLOW}(\beta)$-convergent and F_n is $\texttt{UNS}(\gamma)$-unstable,

$$\|\theta_n^{t+1} - \theta^\star\| \leq \frac{1}{t^\beta} + t \cdot \frac{\varepsilon(n,\delta)}{[\varepsilon(n,\delta)]^{\nu_\star|\gamma|}} \precsim [\varepsilon(n,\delta)]^{\nu_\star},$$

when $t \succsim \varepsilon(n,\delta)^{-\frac{1}{1+\beta}}$

Example of unstable and slow operators

- $\{(X_i, Y_i)\}_{i=1}^n$ are generated from a noisy non-linear regression model of the form

  $$Y_i = (X_i\theta^\star)^2 + \xi_i, \qquad \text{for } i = 1, \ldots, n$$

  where $\theta^\star = 0$
- $\xi_i \sim \mathcal{N}(0, 1)$ and $X_i \sim \mathcal{N}(0, 1)$
- We apply cubic-regularized Newton's method (CNM) to the least-squares loss

Example of unstable and slow operators

- $\widetilde{\mathcal{L}}$ and $\widetilde{\mathcal{L}}_n$ are population and sample least-square losses
- Population CNM operator:

$$\begin{aligned}\mathrm{F}^{\mathrm{CNM}}(\theta) &= \arg\min_{y \in \mathbb{R}} \left\{ \widetilde{\mathcal{L}}'(\theta)(y-\theta) + \frac{1}{2}\widetilde{\mathcal{L}}''(\theta)(y-\theta)^2 + L\left|y-\theta\right|^3 \right\} \\ &= \theta - \frac{\frac{2}{3}\theta^3}{\theta^2 + \sqrt{\theta^4 + \frac{2}{3}\theta^3}}\end{aligned}$$

$\Longrightarrow \mathrm{F}^{\mathrm{CNM}}$ is $\texttt{SLOW}(2)$-convergent

- Sample CNM operator:

$$\mathrm{F}_n^{\mathrm{CNM}}(\theta) = \arg\min_{y \in \mathbb{R}} \left\{ \widetilde{\mathcal{L}}_n'(\theta)(y-\theta) + \frac{1}{2}\widetilde{\mathcal{L}}_n''(\theta)(y-\theta)^2 + L\left|y-\theta\right|^3 \right\}$$

$\Longrightarrow \mathrm{F}_n^{\mathrm{CNM}}$ is $\texttt{UNS}(-\frac{1}{2})$-unstable over the annulus $\mathbb{A}(\theta^\star, \widetilde{\rho}_n, 1)$ with $\widetilde{\rho}_n \asymp \log(n/\delta)/n^{1/4}$

Example of unstable and slow operators

- $\mathrm{F}^{\mathrm{CNM}}$ is $\texttt{SLOW}(2)$-convergent
- $\mathrm{F}_n^{\mathrm{CNM}}$ is $\texttt{UNS}(-\frac{1}{2})$-unstable over the annulus $\mathbb{A}(\theta^\star, \widetilde{\rho}_n, 1)$ with $\widetilde{\rho}_n \asymp \log(n/\delta)/n^{1/4}$
- Additional regularity condition:

$$\left|\mathrm{F}_n^{\mathrm{CNM}}(\theta)\right| \geq |\theta_n^*|$$

for all $|\theta| \in [|\theta_n^*|, 1]$ where θ_n^* is global solution of least-squares loss

- Under suitable initialization, the sequence $\theta_n^{t+1} = \mathrm{F}_n^{\mathrm{CNM}}(\theta_n^t)$ satisfies

$$|\theta_n^t - \theta^\star| \precsim n^{-1/4} \qquad \text{when } t \succsim n^{1/6}$$

Summary of results

Operator Properties	Optimization Rate	Stability	Iterations for convergence	Statistical error on convergence
General expressions				
Fast, stable	$\texttt{FAST}(\kappa)$	$\texttt{STA}(\gamma)$	$\log(1/\varepsilon(n,\delta))$	$\varepsilon(n,\delta)$
Slow, stable	$\texttt{SLOW}(\beta)$	$\texttt{STA}(\gamma)$	$\varepsilon(n,\delta)^{-\frac{1}{1+\beta-\gamma\beta}}$	$[\varepsilon(n,\delta)]^{\frac{\beta}{1+\beta-\gamma\beta}}$
Fast, unstable	$\texttt{FAST}(\kappa)$	$\texttt{UNS}(\gamma)$	$\log(1/\varepsilon(n,\delta))$	$[\varepsilon(n,\delta)]^{\frac{1}{1+\lvert\gamma\rvert}}$
Slow, unstable	$\texttt{SLOW}(\beta)$	$\texttt{UNS}(\gamma)$	$[\varepsilon(n,\delta)]^{-\frac{1}{1+\beta}}$	$[\varepsilon(n,\delta)]^{\frac{\beta}{1+\beta+\lvert\gamma\rvert\beta}}$
Examples				
Fast, stable	$e^{-\kappa t}$	$\frac{r}{\sqrt{n}}$	$\log n$	$n^{-1/2}$
Slow, stable	$\frac{1}{\sqrt{t}}$	$\frac{r}{\sqrt{n}}$	$n^{1/2}$	$n^{-1/4}$
Fast, unstable	$e^{-\kappa t}$	$\frac{1}{r\sqrt{n}}$	$\log n$	$n^{-1/4}$
Slow, unstable	$\frac{1}{t^2}$	$\frac{1}{\sqrt{r}\sqrt{n}}$	$n^{1/6}$	$n^{-1/4}$